Discrete Fréchet Distance for Uncertain Points

Maike Buchin* Stef Sijben*

Abstract

We consider the problem of computing the discrete Fréchet distance between polygonal curves with uncertain points, i.e. the coordinates of the vertices are not known exactly, but are given by a probability distribution. In this case, the discrete Fréchet distance is a random variable. We show that the distribution function for a given coupling can be efficiently evaluated and give an algorithm to compute the coupling with maximum probability of realizing a given discrete Fréchet distance.

1 Introduction

The discrete Fréchet distance is a popular measure for the similarity of two polygonal curves with many applications. In the standard case where the locations of the vertices are known exactly, it can be computed in $O(n^2)$ time for two curves of length at most n [5]. Recently, Agarwal et al. discovered a subquadratic time algorithm for this problem [1].

In practice, the curves are often based on trajectories collected from a moving entity, e.g. using a GPS tracking device. These devices do not provide precise locations, but rather an estimate of the location, typically including an error margin. If the goal is to compute the discrete Fréchet distance, a single outlying observation may lead to a very different coupling than the distance based on real locations, as is illustrated in Figure 1. One proposed solution is the (discrete) Fréchet distance with shortcuts [4, 3], which can remove outliers from one of the curves. Here we will follow the approach of including the uncertainty in the model.

Several models have been proposed to incorporate uncertainty in geometric problems [7]: In the *imprecise points* model each point lies in a given region. With *indecisive points*, each point is selected from a finite set of candidate locations. For *uncertain points*, the location of a point is described using a probability distribution based on the observed location. The uncertain points model is the most general, and it is closest to the practical applications, where a point is likely to be close to the observed location, but large errors are possible. The most commonly used distribution is a circular normal distribution with the mean

*Department of Mathematics, Ruhr-Universität Bochum, {Maike.Buchin,Stef.Sijben}@rub.de

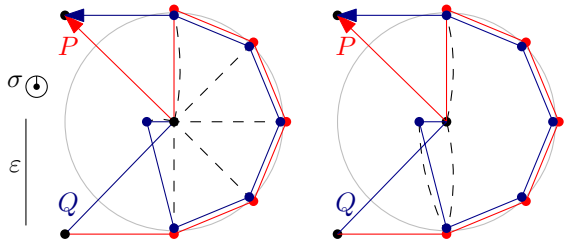

Figure 1: Example of two trajectories (slightly shifted for visual clarity) with discrete Fréchet distance ε. With precise points, all points on the circle are matched to a point in the middle (left). With uncertain points (all normally distributed with standard deviation σ), the optimal coupling matches most vertices to one centered at the same location (right).

at the observed location. The variance can be given by the tracking device's estimated error or based on other measurements.

In the case of imprecise points, where the points are known to be in a given region, efficient algorithms exist to compute the smallest possible discrete Fréchet distance if the regions are d-dimensional balls or for axis-parallel boxes under the L_∞ norm [2]. Computing an upper bound for the discrete Fréchet distance is NP-hard in the imprecise setting [6].

2 Preliminaries

Formally, a polygonal curve with uncertain points P is a sequence of vertices $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_n$. $\boldsymbol{P}_i$ is a random point in $\mathbb{R}^d$ distributed according to a certain (known) distribution. For example, a vertex may be obtained by a GPS fix at a certain time and be normally distributed around the observed location with a certain variance: $\boldsymbol{P}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2)$. We assume that all $\boldsymbol{P}_i$ are independent.

Consider two polygonal curves with uncertain points P and Q of length n and m respectively and assume w.l.o.g. that $n \geq m$. A *coupling* C between P and Q is a sequence of pairs $(a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)$ such that $a_1 = b_1 = 1$, $a_k = n$, $b_k = m$ and for each $i \in \{1, \ldots, k-1\}$ one of the following holds:

- $a_{i+1} = a_i$ and $b_{i+1} = b_i + 1$,
- $a_{i+1} = a_i + 1$ and $b_{i+1} = b_i$, or
- $a_{i+1} = a_i + 1$ and $b_{i+1} = b_i + 1$.

This is an extended abstract of a presentation given at EuroCG 2016. It has been made public for the benefit of the community and should be considered a preprint rather than a formally reviewed paper. Thus, this work is expected to appear in a conference with formal proceedings and/or in a journal.

The discrete Fréchet distance of polygonal curves P and Q is defined as

$$d_{dF}(P,Q) = \min_{C} \max_{i\in\{1,\dots,k\}} |\boldsymbol{P}_{a_i} - \boldsymbol{Q}_{b_i}|,$$

where C ranges over all couplings between P and Q.

The discrete Fréchet distance is usually computed using the free space matrix: A table of size $n \cdot m$, where each cell (i,j) represents a pair of points $\boldsymbol{p}_i, \boldsymbol{q}_j$ from P and Q, respectively. This cell is free if $|\boldsymbol{p}_i - \boldsymbol{q}_j| \leq \varepsilon$. Then, $d_{dF}(P,Q) \leq \varepsilon$ if and only if there is a bimonotone path from $(1,1)$ to (n,m) visiting only free cells.

If the distribution has unbounded support, there are no upper or lower bounds on the value of the discrete Fréchet distance (other than the trivial lower bound of 0). Instead, its value is given by a probability distribution. Ideally, one would like to be able to compute properties of this distribution, e.g.

- distribution function $F(\varepsilon) = \mathbb{P}\left[d_{dF}(P,Q) \leq \varepsilon\right]$,
- probability density $f(\varepsilon) = \frac{d}{d\varepsilon}F(\varepsilon)$,
- quantiles $F^{-1}(\rho) = \inf\{\varepsilon \in \mathbb{R}^{\geq 0} \mid \rho \leq F(\varepsilon)\}$.

Note that if an algorithm for the distribution function is known, the others can be approximated using standard numerical techniques.

Figure 1 shows a pair of trajectories where uncertain points lead to a different result than precise points, in the sense that the coupling with the highest probability of achieving discrete Fréchet distance $\leq \varepsilon$ is different from the coupling with precise points at distance ε. For noisy data, the coupling produced using uncertain points seems more reasonable. In practice, e.g. when studying trajectories with GPS error, one is often not so much interested in the exact value of ε, but in finding a reasonable value for ε with a coupling that provides a good match between the two trajectories. In some cases, using uncertain points avoids intuitively unappealing couplings in favour of a more reasonable one that has a slightly larger distance.

One option to compute $F(\varepsilon)$ is to fix the position of each point, test whether $d_{dF}(P,Q) \leq \varepsilon$ for these locations and integrate each point over $\mathbb{R}^d$, weighted by the probability density of the point, i.e.:

$$F(\varepsilon) = \int_{\mathbb{R}^d} f_{\boldsymbol{P}_1}(\boldsymbol{p}_1) \dots \int_{\mathbb{R}^d} f_{\boldsymbol{P}_n}(\boldsymbol{p}_n) \int_{\mathbb{R}^d} f_{\boldsymbol{Q}_1}(\boldsymbol{q}_1) \dots \int_{\mathbb{R}^d} f_{\boldsymbol{Q}_m}(\boldsymbol{q}_m) I(\varepsilon) \, d\boldsymbol{q}_m \dots d\boldsymbol{q}_1 d\boldsymbol{p}_n \dots d\boldsymbol{p}_1,$$

where $I(\varepsilon)$ is the indicator function which is 1 if $d_{dF}(P,Q) \leq \varepsilon$ for the given coordinates and 0 otherwise.

However, evaluating these $d(n+m)$ nested integrals to a reasonable precision requires time exponential in the dimension [8], so other approaches are needed.

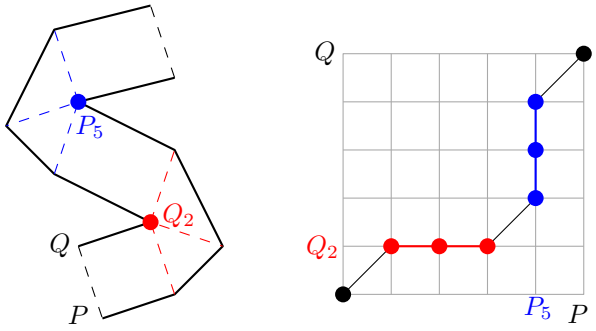

Figure 2: Two trajectories with a coupling and the corresponding free space matrix. The distribution functions for the red and blue segments, as well as each of the black points, can be computed independently.

3 Evaluating $F(\varepsilon)$ for a fixed coupling

We first consider how to evaluate $F_C(\varepsilon)$, i.e. the probability that a given coupling C realizes $d_{dF}(P,Q) \leq \varepsilon$. C can be represented by a bimonotone path through the free space matrix. We assume that this path makes no 90° turns. Any path can be converted to this form in linear time by diagonally going from the cell before the turn to the cell after the turn. This transformation can only increase $F_C(\varepsilon)$, since it removes some diagonals in the coupling and does not introduce any new ones.

A diagonal move from (i,j) to $(i+1,j+1)$ in the path breaks the trajectories up into two subtrajectories each such that

$$d_{dF}(P,Q) = \max\{d_{dF}(P[1\dots i], Q[1\dots j]), d_{dF}(P[i+1\dots n], Q[j+1\dots m])\}.$$

Since these are disjoint subtrajectories, the distributions of their discrete Fréchet distances are independent. This property allows us to break the path up into (possibly degenerate) horizontal and vertical segments which can be treated independently.

Let $C_1, \dots, C_k$ be the segments of the coupling defined above and let $F_{C_\ell}(\varepsilon)$ be the probability that the segment C_ℓ realizes a discrete Fréchet distance between the induced subtrajectories of at most ε. We use independence to obtain $F_C(\varepsilon)$:

$$F_C(\varepsilon) = \prod_{\ell=1}^{k} F_{C_\ell}(\varepsilon).$$

As illustrated in Figure 2, there are three possible cases for $F_{C_\ell}(\varepsilon)$, which are easy to deal with:

1. The segment contains only a single point. This represents one point on each trajectory being matched to each other, thus the question reduces to $F_{C_\ell}(\varepsilon) = \mathbb{P}\left[|\boldsymbol{P}_i - \boldsymbol{Q}_j| \leq \varepsilon\right]$.

For normally distributed points the distance is related to the noncentral chi-squared distribution and this can be computed by evaluating its distribution function once.

2. A vertical segment represents a single point $\boldsymbol{P}_i$ being matched to several points $Q[j \dots j']$. Here, we fix $\boldsymbol{P}_i$, compute the probability that all other points are close enough to the fixed point and integrate $\boldsymbol{P}_i$ over all possible locations:

$$F_{C_\ell}(\varepsilon) = \int_{\mathbb{R}^d} f_{\boldsymbol{P}_i}(\boldsymbol{x}) \prod_{k=j}^{j'} \mathbb{P}\left[|\boldsymbol{Q}_k - \boldsymbol{x}| \le \varepsilon\right] d\boldsymbol{x}. \tag{1}$$

Evaluating these d nested integrals takes $O(c^d \cdot (j' - j))$ time, where c depends on the number of integration steps required, i.e. the integration technique used, the desired precision and details of the functions being integrated [8].

Again, for normally distributed points the probability can be computed using the distribution function of the noncentral chi-squared distribution.

3. A horizontal segment can be processed symmetrically to case 2.

All $F_{C_\ell}(\varepsilon)$ and hence $F_C(\varepsilon)$ can be computed in $O(c^d \cdot (n + m))$ time, where c again depends on the number of integration steps in cases 2 and 3.

4 Computing the optimal coupling

In this section, we present a dynamic programming algorithm that computes an optimal coupling for curves P and Q and distance ε, that is a coupling C for which $F_C(\varepsilon)$ is maximal, i.e. has the highest probability of achieving discrete Fréchet distance at most ε. The probability reached by this coupling is a lower bound for the distribution function $F(\varepsilon)$. Observe that the optimal coupling is one of the form described before, i.e. without 90° turns in the free space.

We use a table similar to the free space matrix, in which each cell represents a prefix of each trajectory and a path from the lower left corner to the cell represents a partial coupling. Using the independence of path segments separated by a diagonal move, we can decompose the optimal path to cell (i, j) into a final horizontal or vertical segment C_k from (i', j') to (i, j) (with $i' \le i$, $j' \le j$ and either $i' = i$ or $j' = j$), and an optimal path to $(i' - 1, j' - 1)$. These paths are then connected using a diagonal edge.

Let $p(i, j)$ denote the probability that the optimal coupling ending at (i, j) realizes $d_{dF}(\tau_1[1 \dots i], \tau_2[1 \dots j]) \le \varepsilon$ and let $\pi(i, j)$ be the coordinates (i', j') where this final segment starts. If the final segment is vertical and (i', j') is known, the probability is given by

$$\begin{aligned} p_v(i, j) &= p(i' - 1, j' - 1) \cdot F_{C_k}(\varepsilon) \\ &= p(i' - 1, j' - 1) \cdot \mathbb{P}\left[\varepsilon \ge \max_{k \in \{j', \dots, j\}} |\boldsymbol{P}_i - \boldsymbol{Q}_k|\right] \\ &= p(i' - 1, j' - 1) \\ &\qquad \cdot \int_{\mathbb{R}^d} f_{\boldsymbol{P}_i}(\boldsymbol{x}) \prod_{k=j'}^{j} \mathbb{P}\left[|\boldsymbol{Q}_k - \boldsymbol{x}| \le \varepsilon\right] d\boldsymbol{x}. \end{aligned}$$

If the final segment is horizontal, a similar expression exists for $p_h(i, j)$ and $p(i, j) = \max\{p_v(i, j), p_h(i, j)\}$.

To find (i', j') we search all possible predecessors, i.e. all cells in the jth row or ith column preceding (i, j), including (i, j) itself, select the (i', j') that maximizes the probability and set $p(i, j)$ and $\pi(i, j)$ accordingly. Then we construct the optimal coupling by following the $\pi(i, j)$ pointers back from (n, m).

The table contains $O(n^2)$ cells, for each cell we need to test $O(n)$ predecessors and computing $p(i, j)$ for a fixed predecessor takes $O(c^d \cdot (n))$ time to evaluate the integral as discussed before.

Theorem 1 *Given two curves with uncertain points P and Q of length n and a threshold ε, the coupling that has maximum probability of realizing $d_{dF}(P, Q) \le \varepsilon$ can be computed in time $O(c^d n^4)$, where c depends on the number of integration steps.*

Instead of a single coupling, the best k couplings can be computed by replacing $p(i, j)$ and $\pi(i, j)$ by lists of length k. The running time increases by a factor k.

Note that for fixed (i, j), $F_{C_k}(\varepsilon)$ is an increasing function in i' and j', since fewer points are matched to the fixed point as the segment becomes shorter. Thus, better running times are achieved in practice by searching predecessors backward from (i, j), stopping the search when $F_{C_k}(\varepsilon)$ for some (i', j') becomes smaller than the best known lower bound for $p(i, j)$.

Instead of computing the optimal coupling for a fixed distance ε, given a fixed probability ρ we can compute the coupling C that realizes $F_C(\varepsilon) \ge \rho$ for the smallest value of ε among all couplings, by searching over the distance values, using the dynamic programming algorithm in each step.

5 Experiments

The algorithms described in the previous sections were implemented in R and evaluated for several inputs. For the trajectories and ε shown in Figure 1, these experiments confirm that the coupling shown on the right indeed has a much higher probability (0.14) of realizing $d_{dF}(P, Q) \le \varepsilon$ than the coupling used to realize the discrete Fréchet distance for precise points

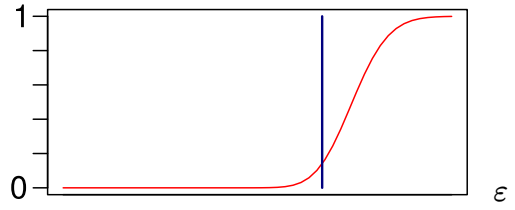

Figure 3: Probability that optimal coupling realizes $d_{dF}(P,Q) \leq \varepsilon$ for the trajectories in Figure 1. The discrete Fréchet distance for precise points is indicated by the blue line.

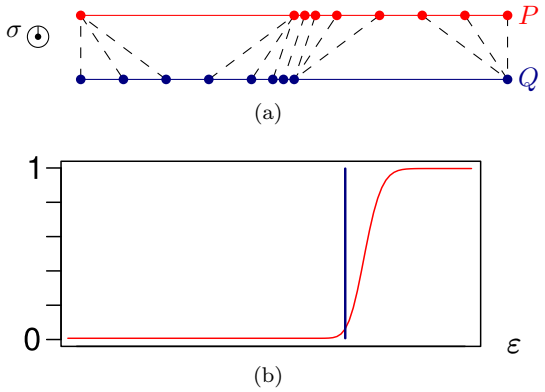

Figure 4: (a) Two curves with the coupling that realizes the minimal $d_{dF}(P,Q)$. Points are normally distributed with standard deviation σ. (b) Probability that the optimal curve realizes $d_{dF}(P,Q) \leq \varepsilon$. The discrete Fréchet distance for precise points is indicated by the blue line.

(0.00057). The probability that the optimal coupling realizes $d_{dF}(P,Q) \leq \varepsilon$ is plotted against ε in Figure 3. The discrete Fréchet distance with uncertain points tends toward slightly larger values than with precise points. This is to be expected, since the discrete Fréchet distance is a bottleneck distance and a single outlying point can cause the distance to become larger. The same can be observed for a different set of curves in Figure 4.

The integration method used is crucial for the accuracy of the algorithm. The expression in Equation 1 is strongly peaked near the mean locations of the input points, and the function must be sampled sufficiently densely near these locations. Our implementation uses the R package `cubature`[1].

The running time of the algorithm depends highly on the input trajectories and ε. The reason for this is that the heuristic described at the end of Section 4 was implemented, which in some cases terminates the predecessor search much earlier than in the worst case. In general the running time is lower for small ε than for large values. For the examples shown, with $n = 9$, the running time can be up to 1 minute.

[1] https://cran.r-project.org/web/packages/cubature/

6 Conclusion

We discussed the problem of computing the discrete Fréchet distance between polygonal curves with uncertain points. Many interesting questions remain open. The main question is whether the distribution function $F(\varepsilon)$ can be computed efficiently. Another direction for future work is improving the running time of the algorithm presented, either in general or for specific classes of trajectories.

References

[1] P. K. Agarwal, R. B. Avraham, H. Kaplan, and M. Sharir. Computing the discrete Fréchet distance in subquadratic time. *SIAM Journal on Computing*, 43(2):429–449, 2014.

[2] H.-K. Ahn, C. Knauer, M. Scherfenberg, L. Schlipf, and A. Vigneron. Computing the discrete Fréchet distance with imprecise input. *International Journal of Computational Geometry & Applications*, 22(01):27–44, 2012.

[3] R. B. Avraham, O. Filtser, H. Kaplan, M. J. Katz, and M. Sharir. The discrete Fréchet distance with shortcuts via approximate distance counting and selection. In *Proceedings of the thirtieth annual symposium on Computational geometry*, page 377. ACM, 2014.

[4] A. Driemel and S. Har-Peled. Jaywalking your dog: computing the Fréchet distance with shortcuts. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 318–337. SIAM, 2012.

[5] T. Eiter and H. Mannila. Computing discrete Fréchet distance. Technical report, Technische Universität Wien, 1994.

[6] C. Fan and B. Zhu. Complexity and algorithms for the discrete Fréchet distance upper bound with imprecise input. *arXiv preprint arXiv:1509.02576*, 2015.

[7] A. Jorgensen, M. Löffler, and J. M. Phillips. Geometric computations on indecisive and uncertain points. *arXiv preprint arXiv:1205.0273*, 2012.

[8] W. H. Press, S. A. Teukolsky, T. Veterling, and B. Flannery. *Numerical recipes: The art of scientific computing*, pages 196–200. Cambridge university press, 3rd edition, 2007.